午前3時のデータマイニング

第2回：三河屋サブちゃんとマイニングする!?

奥山真一郎
OKUYAMA Shin-ichiro


illustration by T. Chikatsu

♪♪うーらのはーたけでポチがなくー。正直じーさん掘ったらばー、おーばーあんこーばあんが、ざーっくざーっくざっくざく♪♪　皆さん、おはようございます。「午前3時のデータマイニング」の時間です。先月に引き続き、今月も楽しくデータマイニングを学んでまいりましょう。今回は、データマイニングの理解をさらに深めるため、データマイニングと既存の分析手法との違いを見ながら、データマイニングの特徴と実践の方向性を考えていくことにしましょう。

三河屋サブちゃんとクロスセリングを考える

前回は、データマイニングとは何かについてその概要を見てきました。あれから1ヶ月経ちましたが、皆さん覚えていますか？ マイニングの応用例として「クロスセリング」という言葉が出てきましたね。何のこと？ と忘れてしまった方のために、まずは簡単に復習をしておきましょう。

クロスセリングとは、いわゆる「合わせ売り」のことです。ここで三河屋サブちゃんの例を考えてみます。皆さん、三河屋のサブちゃんってご存知ですか？ そうです。あの「サザエさん」に登場する、磯野家御用達の酒屋「三河屋」のお兄さんです。いつも元気に磯野家の勝手口から「ちわー！ 三河屋でーす！」と注文のあった品物の配達にやってきます。時には、御用聞きにうかがうという、まさに顧客密接型のマーケティング戦略を展開しています。

例えば、サブちゃんが磯野家から注文を受け、「みりん」を配達にきたとします。すると、サブちゃんはすかさず、サザエさんに対して「毎度ありがとうございます。他にご用はございませんか？」との決まり文句に続けて、「いやー。今日は新潟の蔵元からいいお酒が入っていますよ！」などとさりげなく購買を促します。これが、いわゆる「クロスセリング」です。

ここで問題になるのは、どの商品を選んでサザエさんに薦めるか？ ということです。サザエさんは、大変"賢い"(?)消費者の1人ですので、単純に今日売りたい商品をただ紹介しただけでは「今日は結構よ！ おほほほほ」と軽くあしらわれてしまうでしょう。また、あまりニーズのない商品をしつこく薦めても、逆にうっとおしがられるのが人情です。

やはり，ここでは「そろそろ醤油が切れていませんか？」とか「マスオさんにぴったりのワインが入りましたよ」など、タイムリーかつお客様のニーズに合った情報を提供して磯野家を攻めたいところです。しかし、サブちゃんはいささか先生のお宅など、他にもたくさんのお得意様を抱えています。すべてのニーズを把握するのは大変ですね。

そこで、データマイニングが登場するわけです。その方法はいろいろ考えられますが、まずは取引の履歴データを活用し、従来の配達作業からプラスアルファのビジネスチャンスを拓くことを考えてみましょう。

検索・集計を超えろ！

取引データとは、表のような「いつ」「誰が」「何を」「いくらで」「どれくらい」買ったかというようなシンプルなデータです。ここでは、個人情報ではなく世帯情報という視点から見ることにしましょう。

ただし、ここで気を付けなければならないのは、前回でも触れたように データをやみくもにこねくり回しても仕方がないということです。データを分析する前には、必ず「何を知りたいのか」「どのようにすれば分かるのか」を準備しなければなりません。さもなければ、たちまち「データジャングル」の中に迷い込んでしまいます。では、何を知ればサブちゃんはハッピーに（販売数を伸ばせる）なれるのでしょうか？

表 「データレギュラ・ト」[illegible]

顧客ID	日付	購入品目	点数	単価	金額
0000001	1999/10/1	ビール	12	350	4200
0000001	1999/10/2	みりん	1	400	400
0000001	1999/10/2	焼酎	1	300	300
0000002	1999/10/1	ビール	6	350	2100
0000002	1999/10/1	赤ワイン	1	1500	1500
0000003	1999/10/1	焼酎	1	300	300
0000003	1999/10/1	ビール	6	350	2100
0000003	1999/10/2	日本酒	1	1800	1800
0000003	1999/10/2	しょうゆ	1	400	400
0000004	1999/10/1	ビール	5	350	1750
0000004	1999/10/1	赤ワイン	1	1500	1500
0000004	1999/10/2	ビール	24	350	8400
0000004	1999/10/2	焼酎	1	300	300

●RFMを軸に顧客の総合指標化

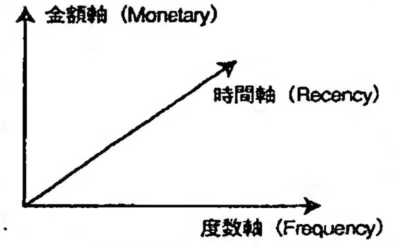


図1 RFM分析

- ●「最近、商品を多く買ってくれるお客さんは誰？逆に、最近ご無沙汰なお客さんは？」
- ●「いつも（回数的に）ちょくちょく買ってくれるお客さんは誰？」
- ●「いつも（金額的に）たくさん買ってくれるお客さんは誰？」

[illegible]

これらの質問は、[illegible]ることができます。

- ・時間軸(Recency)
- ・度数軸(Frequency)
- ・金銭軸(Monetary)

[illegible]

●「月に5万円以上のお買い上げのあるお客さんの中で、一番売れている商品は何？」

[illegible]

●「月10万円以上の商品を購入したお客さんの中で、一番多い商品の組み合わせは？」

[illegible]

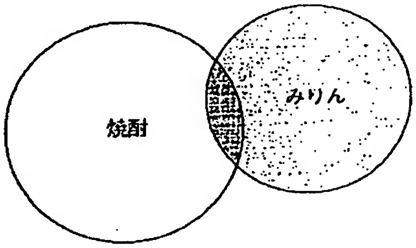
検索の処理だけでは、簡単に答えを出すことができません。しかし、この情報をつかめば、お得意様にお薦めする商品が浮かび上がってきますね。ではどうすればよいのでしょうか？

これぞマイニングの出番です。あらゆる組み合わせを探索し、件数の多い順にレポートする、ただそれだけの作業のような気がしますが、商品の組み合わせは２品目に限ったことではありません。取り扱い商品の中からすべての組み合わせを取り出すとなると、組み合わせの数は膨大になり、その中から有効な組み合わせを見つけるのは大変なことです。

データマイニングでは、この組み合わせの抽出を「アソシエーションルール」(Association Rule)と呼びます。アソシエーションは、**図２**のように、サポート(Support)とコンフィデンス(Confidence)という２つの基準により、特徴的な組み合わせを評価するものです。

サポートとは、全顧客(ユニークなIDの数)の中で、その商品の組み合わせで購入した人が何人(ここでは何世帯)いるのかを表わした指標です。例えば、「焼酎」と「みりん」の組み合わせで購入した世帯数は、全体の何パーセントになるのか、というようなものです。

数値例：
全顧客世帯数＝100世帯
焼酎かつみりん購買世帯数＝20世帯
焼酎購入数＝50世帯
みりん購入世帯数＝40世帯



焼酎購入世帯が同時に、みりんも購入する→20/50＝40%（コンフィデンス）
全顧客世帯の中で焼酎購入者が、みりんも購入する→20/100＝20%（サポート）

図２：アソシエーションはサポートとコンフィデンスにより抽出できる

これに対してコンフィデンスとは、ある商品を購入した顧客の中で、同時に別の商品と組み合わせて購入する割合を言います。例えば、「焼酎」を購入した世帯の中で、「みりん」も同時に購入した世帯の割合を指します。

これらの方法によって求められた組み合わせの中から、びっくりするようなおもしろいルールを探し出せることがデータマイニングの魅力の１つといえるでしょう。ここで紹介したアソシエーションルールの抽出は、データマイニングの最も特徴的な手法であり、非常に簡単なものですのでいろいろ応用できます。例えば、前回紹介したビールとおむつの関連も、このサポートとコンフィデンスにより浮かび上がってきたアソシエーションルール、「ビール→（ならば）おむつ」です。

OLAPとの違いは何か

いかがでしたか？磯野家攻略の糸口と、マイニングの特徴が浮かび上がってきたでしょうか。では、ここでデータマイニングと比較すべきデータ分析のアプローチについてちょっと触れておきましょう。

皆さん、OLAPってご存知ですか？意外にOLAP(オーラップと読みます)とデータマイニングを混同されている方が多いようなので、今一度整理しておきましょう。OLAPとは、On-Line Analytical Processingの略で、リレーショナルデータベース(RDBMS)の弱点を補完するために、米国のコッド博士が提唱した12のルールに基づいて発展した分析手法です(**図３**)。

日本では、「多次元データ分析」と訳されていますが、データ解析の手法としての視点からいえば、実はOLAPはデータの「集計」にすぎません。多次元データ分析と聞くと、なんとなくすごそうな感じがしますが、実際には「オンライン集計処理」「集計済み結果表示」の仕組みとして分類できます。集計結果は、クロス集計のイメージやグラフで表現できます。

当然ですが、オンライン上でクロス集計結果を見ることができるといっても、１つ１つのデータを眺めるのは人間の役割です。サザエさんの例では、サブちゃん自身があらゆるクロス集計表を眺めながら考える姿をイメージすれば、雰囲気が分かると思います。

また、OLAPでは紙にプリントアウトしたクロス集計表とは異なり、ドリ

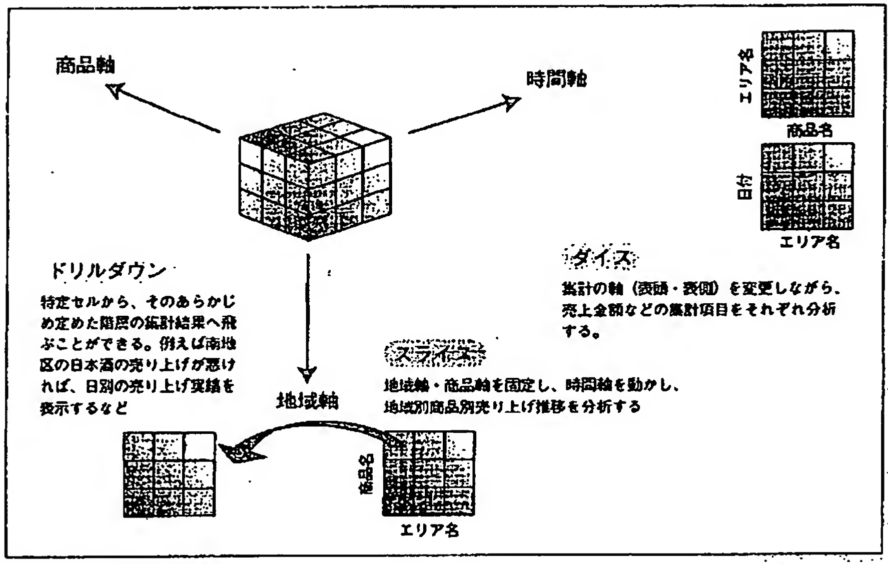


図3：OLAP の機能概要

ルダウン(Drill Down)やスライス＆ダイス(Slice & Dice)という機能により、必要な集計項目を自由に変えたり、あらかじめ定めた集計の次元を切り替えながら結果を見ることができます。必要な項目を定めれば、常に結果をモニタリングできるので、大変便利な機能と言えるでしょう。

しかし、私たちはおもちゃのルービックキューブでさえ、組み変えているうちに元に戻せなくなってしまいます。データ探索にOLAPを用いるということは、実質的に集計・検索を繰り返すことと同じです。このため、データのジャングルに迷い込んでしまう可能性も高いと言えるでしょう。

これに対してマイニングでは、先のアソシエーションのように、注目すべき組み合わせをレポートしてくれます。探索そのものにかける時間を短縮できるため、その結果を考察するという本来の業務に多くの時間を費やすことができるのです。

マイニングから得られた結果は、OLAPの次元として採用することも可能です。また、ドリルスルーといった詳細データをクロス表の特定セルから参照するような機能と組み合わせることもできます。マイニングとOLAPは、目的に合わせてそれぞれを使い分けていくことが大切です。

たかがアソシエーション、されどアソシエーション

今月はデータマイニングの特徴を三河屋のサブちゃんと一緒に考えてきました。アソシエーションルール、クロスセリングについては、先月に引き続き細かく見てきましたので、お分かりいただけたと思います。

このほかにもクロスセリングは、商品の組み合わせだけでなく、顧客属性や天気など、さまざまな外部情報も組み込むことができます。例えば、「男性 & 30代 & ライトビール → 赤ワイン」など、より複雑な連関ルールを探索することが可能となるわけです。そして、このようなアソシエーションルールの発見によるクロスセリングは、小売業だけで応用されているものではありません。例えば、不良債権の原因を探索するなど、金融商品の分析もできますし、Webのアクセスログを利用すれば、サイトとサイト、ページとページの連関を発見することができます。また、医薬品の市販後の調査など、副作用をもたらす処方の組み合わせの探索もできるでしょう。

このように、アソシエーションにデータをどのように与えていくかのアイデア1つによって、いろいろなことが見えてきます。埋もれた大量のデータにはまだまだダイヤモンドが眠っているのです。ぜひ、皆さんもアソシエーションの身近な応用例を考えてみてください。きっと、サブちゃんも喜びますよ！ どうやら今月は、裏の畑で「たま」も鳴いているようですね。 DBM

奥山真一郎（おくやましんいちろう）
国際大学修了後、SASインスティチュートジャパン入社。データサイエンスグループにて、統計解析、データマイニングのコンサルテーションに従事する。現在は「孤高のDataMiner」として、データとの付き合い方を模索中。
E-Mail：FZH04331@nifty.ne.jp